UZH@CRAFT-ST: a Sequence-labeling Approach to Concept Recognition

Furrer, Lenz ; Cornelius, Joseph ; Rinaldi, Fabio

Abstract: As our submission to the CRAFT shared task 2019, we present two neural approaches to concept recognition. We propose two different systems for joint named entity recognition (NER) and normalization (NEN), both of which model the task as a sequence labeling problem. Our first system is a BiLSTM network with two separate outputs for NER and NEN trained from scratch, whereas the second system is an instance of BioBERT fine-tuned on the concept-recognition task. We exploit two strategies for extending concept coverage, ontology pretraining and backoff with a dictionary lookup. Our results show that the backoff strategy effectively tackles the problem of unseen concepts, addressing a major limitation of the chosen design. In the cross-system comparison, BioBERT proves to be a strong basis for creating a concept-recognition system, although some entity types are predicted more accurately by the BiLSTM-based system.

DOI: https://doi.org/10.18653/v1/D19-5726

Originally published at:
Furrer, Lenz; Cornelius, Joseph; Rinaldi, Fabio (2019). UZH@CRAFT-ST: a Sequence-labeling Approach to Concept Recognition. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Hong Kong, China, 4 November 2019 - 5 November 2019. Association for Computational Linguistics, 185-195.
DOI: https://doi.org/10.18653/v1/D19-5726

UZH@CRAFT-ST: a Sequence-labeling Approach to Concept Recognition

Lenz Furrer[*†] **Joseph Cornelius**[*] **Fabio Rinaldi**[*†]
[*]University of Zurich, Institute of Computational Linguistics
[†]Swiss Institute of Bioinformatics
Andreasstr. 15, CH-8050 Zürich, Switzerland
{lenz.furrer,joseph.cornelius,fabio.rinaldi}@uzh.ch

Abstract

As our submission to the CRAFT shared task 2019, we present two neural approaches to concept recognition. We propose two different systems for joint named entity recognition (NER) and normalization (NEN), both of which model the task as a sequence labeling problem. Our first system is a BiLSTM network with two separate outputs for NER and NEN trained from scratch, whereas the second system is an instance of BioBERT fine-tuned on the concept-recognition task. We exploit two strategies for extending concept coverage, ontology pretraining and backoff with a dictionary lookup. Our results show that the backoff strategy effectively tackles the problem of unseen concepts, addressing a major limitation of the chosen design. In the cross-system comparison, BioBERT proves to be a strong basis for creating a concept-recognition system, although some entity types are predicted more accurately by the BiLSTM-based system.

1 Introduction

We describe our submission to the CRAFT shared task 2019. We participated in the concept annotation (CA) subtask, which comprises biomedical named entity recognition (NER) and normalization (NEN) for full-text scientific articles. We tested two different neural architectures, a BiLSTM-based network trained from scratch and a transformer system obtained by fine-tuning BioBERT. While NER+NEN tasks have often been approached with a pipeline architecture (NER output passed to NEN as input), we strove for tackling both tasks jointly in a single model.

In essence, we cast the task as a sequence-labeling problem, by directly predicting IDs as symbolic labels. This approach has the obvious drawback that the models will only ever predict IDs that were seen in the training data. In order to account for this limitation, we used different strategies to enrich the systems with information derived from terminology resources, such as ontology pretraining and combination with a rule-based dictionary-lookup system.

The source code of our systems is publicly available at https://github.com/OntoGene/craft-st.

2 Data

The CRAFT corpus (Bada et al., 2012; Cohen et al., 2017) is a collection of 97 full-text articles, of which 30 have been released only in the course of the present shared task. The documents were manually annotated with respect to 10 different entity types, linked to 8 manually curated ontologies of biomedical terminology:

CHEBI: chemicals/small molecules (Chemical Entities of Biological Interest)

CL: cell types (Cell Ontology)

GO_CC: cellular and extracellular components and regions (Gene Ontology)

GO_BP: biological processes (Gene Ontology)

GO_MF: molecular functionalities possessed by genes (Gene Ontology)

MOP: chemical reactions and other molecular processes (Molecular Process Ontology)

NCBITaxon: biological taxa and organisms (NCBI Taxonomy)

PR: proteins, genes, and transcripts (Protein Ontology)

SO: biomacromolecular entities, sequence features (Sequence Ontology)

UBERON: anatomical entities (UBERON)

In addition, the annotations are distributed in an extended variant, i. e. **CHEBI_EXT**, **CL_EXT** etc., resulting in a total of 20 annotation sets. For the extension annotations, the creators of the CRAFT corpus modified the given ontologies in

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 185–195
Hong Kong, China, November 4, 2019.

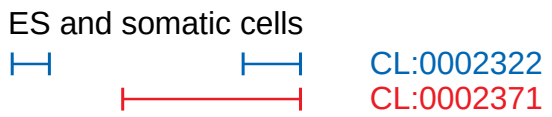

Figure 1: Example of discontinuous and overlapping annotations in an elliptical coordination construction.

a way to better represent actual usage of biomedical entities in scientific texts. In many cases, new concepts were added or existing ones were replaced; some concepts were merged across ontologies (e. g. CL_GO_EXT:cell, which refers to an unspecific cell).

The size of the ontologies varies considerably, ranging from 5 concepts for GO_MF to 1,167,358 concepts for NCBITaxon_EXT. The 67 articles released for training contain a total of 575,296 tokens and the 30 test articles contain 239,409 tokens. In the training set of the corpus, PR_EXT holds the most annotations (19,862 mentions of 1075 unique IDs) and MOP has the fewest (240 mentions of 16 unique IDs). The corpus includes 1264 discontinuous annotations, which are found most frequently among the GO_BP annotations with 493 occurrences. Of these, 788 annotations partially overlap with another annotation of the same type, sharing at least one token (cf. Figure 1).

Furthermore, the corpus contains 3362 annotations that overlap with an annotation of a different type. The three most common combinations are ⟨CL, UBERON⟩ (571), ⟨GO_BP, UBERON⟩ (500) and ⟨CL, GO_BP⟩ (349). The three most common terms with cross-type annotations are "gene expression" (161), "Mcm4/6/7" (107) and "Cln3" (97), whereby the ten most common terms account for 22.159% of the overlapping annotations.

For the present work, we treated each annotation set as a separate dataset independent of all others, resulting in 20 individual tasks. This is in accordance with how the evaluation is carried out.

2.1 Preprocessing

The CRAFT corpus is distributed with annotations in a stand-off format, i. e. separated from the text. The primary format is Knowtator XML, but a format-conversion suite is provided for producing BioNLP format, which is more easily processed and which is also required for the system predictions by the official evaluation suite.

The stand-off formats allow representing interlaced annotations, such as discontinuous spans and overlapping concepts, which often occur together (cf. Figure 1). For sequence classification, however, two parallel sequences of tokens and labels with one-to-one correspondence are required, typically using IOB or IOBES tags. There is no straight-forward method to represent interlaced annotations in this format, even though potential solutions have been proposed (Metke-Jimenez and Karimi, 2016; Dai, 2018). Instead, we decided to use a lossy transformation which simplifies the annotations during the conversion. While this means that our systems cannot represent (and thus predict) all required types of annotations, we believe that the phenomenon is too rare to justify the increase in complexity (multi-class classification for overlaps, additional labels for discontinuity, more complex heuristics in postprocessing).

We used the *standoff2conll* suite[1] for converting the annotations from BioNLP to a CoNLL-like tab-separated format. We chose the "first-span" strategy for resolving discontinuous spans and "keep-longer" for overlapping concepts, the former of which we wrote ourselves in analogy to the existing "last-span" strategy. The *standoff2conll* suite also takes care of sentence splitting and tokenization, using rule-based approaches.

In addition, we applied abbreviation expansion using Ab3P (Sohn et al., 2008). We removed short-form candidates that were all-lowercase, consisted of only one character or had a P-precision (Ab3P's confidence metric) of less than 0.9. For each article, all occurrences of the remaining short forms were then replaced with their best-matching long-form (highest P-precision). Abbreviation expansion was only integrated in the BiLSTM system.

2.2 Postprocessing

Since our systems produce predictions in a CoNLL-like format, an additional conversion step was necessary to meet the requirements of the evaluation suite (BioNLP format). As another contribution to the *standoff2conll* tool, we wrote a converter for the inverted direction (CoNLL to stand-off). The converter is graceful with respect to invalid tag sequences (e. g. O – I – O) and makes use of existing functionality.

[1] https://github.com/spyysalo/standoff2conll

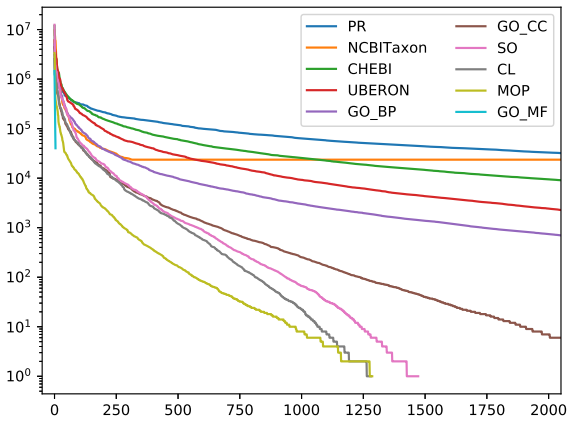

Figure 2: Occurrences of all concepts in the CRAFT ontologies, as annotated by OGER in a large subset of Medline+PMC, sorted by rank.

3 System Description

For the concept annotation task of the CRAFT shared task, we tested two different neural architectures, BiLSTM and transformer (BERT). In addition, we used a rule-based dictionary-lookup system (OGER), which served both as a baseline and as an auxiliary component in the machine-learning systems.

All three systems are applied to each of the annotation sets individually, i. e. each system performs 20 independent predictions. For the neural systems this means that we trained 20 separate models for each configuration; in the case of cross-validation, the number of models is multiplied by another factor.

In a supervised classification setup, an example-based model can only ever predict concepts that have been seen in the training phase. As the concept vocabularies are very large for most of the entity types, an annotated corpus with full coverage is out of reach. However, since the mentions of biomedical concepts resemble a Zipfian distribution (cf. Figure 2), it is often possible to achieve reasonable performance in terms of F-Score even with such a restricted label set. Yet a system that is limited to the concepts of a training corpus is undesirable in many application scenarios. For this reason, we searched for ways to combine the neural systems with the dictionary-based system OGER, which requires no training and can target the entire set of concepts from a given ontology.

Another common challenge of the neural systems, inherent to the sequence-labeling approach, is the classification of multi-word expressions, as each token is labeled individually. This is especially true for semantically weak tokens like stop words, single letters, or numbers (e. g. "I" in "Hexokinase I"). Correctly annotating these tokens is only possible in light of their context, which makes them exceedingly demanding with respect to generalization.

In contrast, OGER annotates multi-word expressions jointly with a single lookup for the entire span. As another difference, OGER can predict multiple concepts for the same span or even interleaved spans, whereas the sequence taggers can only assign one concept to each token.

3.1 Dictionary-based System

OGER (Basaldella et al., 2017; Furrer et al., 2019) is a fast, reliable concept-recognition system based on dictionary lookup. It is highly flexible in terms of matching rules (tokenization, spelling normalization) and supports a wide range of input/output formats. For the present work, we used the following spelling normalization rules: transliteration of Greek letter names, ise/ize conflation, and stemming. Based on the performance on the training set, we fine-tuned the configuration on a per-ontology basis; e. g. stemming was disabled for NCBITaxon and PR.

3.2 BiLSTM-based System

Architecture

Our first neural sequence tagger is a network with a bidirectional Long Short-Term Memory (BiLSTM) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) layer at its core. Its architecture is illustrated in Figure 3. The input tokens x are represented using pretrained word embeddings (Chiu et al., 2016) and randomly initialized character embeddings, the latter of which are transformed into a token-level vector through a convolution and pooling operation (not shown in the figure). The token representation is concatenated with a dictionary feature x^O, which is a vector that encodes the predictions by OGER (using the same dimensionality as the NEN output vector over y^C, see below).

The subsequent layers are inspired by the work of Zhao et al. (2019), who propose a multi-task-learning framework to jointly tackle span detection (NER) and normalization (NEN). A key step to make NER and NEN compatible was to model NEN as a sequence-labeling problem, where IDs

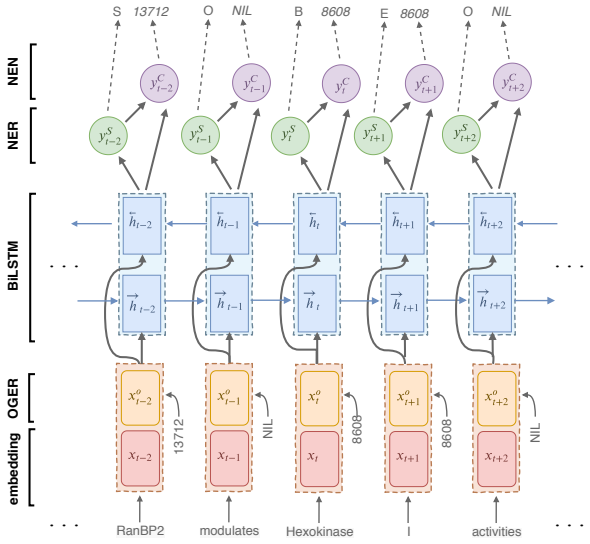

Figure 3: Architecture of the BiLSTM-based sequence tagger (simplified).

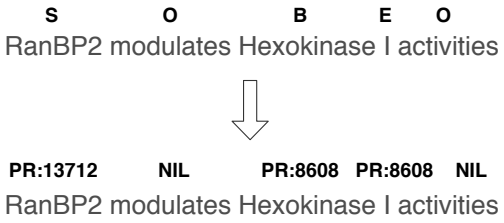

Figure 4: Example of labeling with IOBES tags (NER) and concept IDs (NEN).

are predicted for each token just like span tags in NER (cf. Figure 4). A BiLSTM layer consumes a sequence of token representations one sentence at a time. The sequence representation is then forked into two output layers with soft-max activation, which solve different tasks: The span-detection layer predicts one of the labels $y^S = \{\mathrm{I}, \mathrm{O}, \mathrm{B}, \mathrm{E}, \mathrm{S}\}$, as in a classical single-type NER problem. The normalization layer predicts concept labels (IDs) from $y^C = y^C_T \cup y^C_P \cup y^C_O$, where y^C_T are all labels seen in the training corpus, y^C_P are the labels seen in ontology pretraining and y^C_O are all labels found by OGER. The label set y^C includes the NIL symbol, which denotes the absence of a concept annotation. In addition to the hidden states of the BiLSTM layer, the normalization layer takes the output of the span-detection layer as an input. In contrast to Zhao et al., there is no symmetric feedback between the two output layers, i. e. the span-detection layer does not "see" the output of the normalization layer. This allows training spans and concepts simultaneously.

Training a BiLSTM model for NER and NEN

Training is performed in two phases, ontology pretraining and main training. In the first phase, the model adapts to the domain of the respective entity type by means of terminology entries. At this stage, the model is trained on isolated names and synonyms extracted from the provided ontology files. Due to technical limitations, we restricted the pretraining data to the 1000 most common concepts of each ontology. As an approximation for determining the most commonly used concepts in the literature, we automatically annotated a large subset of Medline (26M abstracts) and PubMed Central (725k articles) with OGER. We sorted the annotated concepts by occurrence and manually removed high-frequency false positives. The model is then pretrained on the top 1000 concepts for a fixed number of 20 epochs.

In the main training phase, training continues with full sentences from the CRAFT corpus. At this stage, the model learns to predict concept mentions in real-world language usage, including contextual hints, frequency distribution, and challenges like rephrasing and non-standard spelling. While the main training is likely to override parts of the connections learnt during ontology pretraining, others may remain to form some kind of background knowledge. Main training is performed as 6-fold cross-validation, where the held-out set of each fold is used to determine when to stop training, using a patience value of 5 epochs. Thus, 6 models are trained for each entity type.

Agreement of NER and NEN Predictions

At prediction time, the softmax scores from all 6 models are averaged before the highest-ranking label for a particular token is determined. Also, when abbreviations have been expanded into multiple tokens during preprocessing, their scores are averaged prior to label selection. The outputs for NER and NEN are tested for agreement. Agreement means that both outputs see a given token t as either relevant or irrelevant, or formally:

$$(\hat{y}^S_t = \mathrm{O} \wedge \hat{y}^C_t = \mathrm{NIL}) \vee (\hat{y}^S_t \neq \mathrm{O} \wedge \hat{y}^C_t \neq \mathrm{NIL})$$

The labels $\hat{y}^S_t$ and $\hat{y}^C_t$ are chosen such that they satisfy the above requirement, while maximizing the overall score. In practice, we compare the score product of the irrelevant labels (O/NIL) to the score product of the top-ranking relevant labels

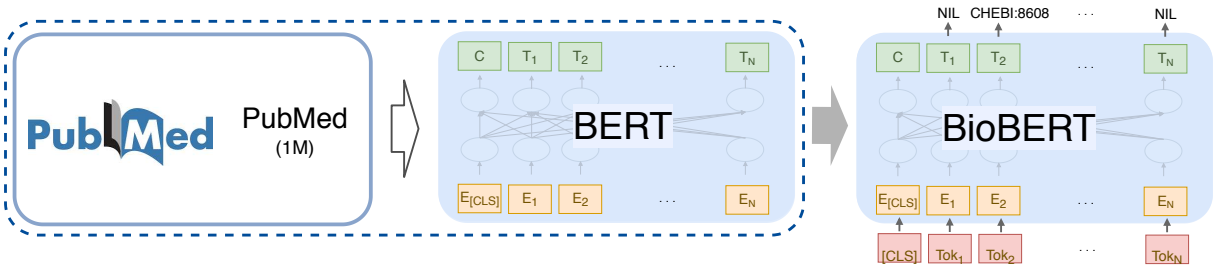

Figure 5: A symbolic illustration of the BioBERT model as a result of a BERT model pretrained on the PubMed corpus and fine-tuned for NEN on the CRAFT corpus.

of either output. This means that we might select a non-best-ranking label for one of the outputs.

3.3 BERT-based System

Background: BERT and BioBERT

The multi-layer BERT model (Devlin et al., 2019) is trained in an unsupervised setting to create bidirectional contextual representations of a token from unlabeled text conditioned on the left and the right context. Two tasks are used to train the BERT model: first, to predict whether two sentences follow each other, and second, to predict a randomly masked token. The resulting pretrained BERT model can be applied to a large number of tasks, such as question answering, next sentence prediction, or NER. Recently it has been shown that the use of pretrained BERT models is especially beneficial to NER tasks (Devlin et al., 2019). In contrast to traditional models used for NER tasks such as long short-term memory (LSTM) models and conditional random field (CRF) models (Habibi et al., 2017), which use context-independent word vector representations such as Word2vec (Mikolov et al., 2013) or GLOVE (Pennington et al., 2014), the BERT model learns context-dependent word vector representations.

A specialized variant of the BERT model for the biomedical domain is the BioBERT (Lee et al., 2019) model, which has been shown to produce state-of-the-art results for NER in the biomedical domain (Jin et al., 2019). The BioBERT model is initialized using the BERT model pretrained on general-domain data (Wikipedia, Bookcorpus) and is then pretrained an additional 200k steps on a corpus of one million PubMed abstracts.

Fine-tuning BioBERT for NER and NEN

For our second system in the CRAFT shared task, we used the readily pretrained BioBERT model available online.[2] We wrote a task-specific head for ID tagging and fine-tuned the model on the CRAFT corpus for another 55 epochs. Like the BiLSTM system, the model is trained to directly predict a sequence of concept IDs from a sequence of input tokens. Technically, we implemented this as an adaptation of an NER tagger by extending the tagset to all concept labels of the training set (cf. Figures 4 and 5).

As a variant, we fine-tuned another BioBERT model as a classical NER tagger over IOBES tags and combined the resulting predictions with annotations from OGER. Predictions were only kept if both OGER and BERT agreed, i. e. both produced a label different from O/NIL. This system, which resembles a traditional NER+NEN pipeline, combines the high recall of the dictionary-based system with the context-aware span detection of an example-based classifier.

Additionally, we combined the previous two systems into a third system. In this variant, the ID tagger takes precedence, whereas the span tagger serves as a backoff model. Whenever the first system does not predict an ID for a token, the backoff system gets a chance to provide an ID, thus joining the forces of two alternative approaches.

3.4 Related Work

Concept-recognition systems solve the task of detecting and linking textual mentions to terminology identifiers. In the past, this problem has often been approached with a pipeline combining an NER tagger with a dictionary-lookup module (e. g. Campos et al., 2013; Ghiasvand and Kate, 2014) or a rule-based system (D'Souza and Ng, 2015; Lee et al., 2016). Leaman et al. (2013) prepared the ground for machine-learning approaches to the normalization task, modeling it as a rank-

[2] https://github.com/naver/biobert-pretrained

		CHE-BI	CL	GO BP	GO CC	GO MF	MOP	NCBI Taxon	PR	SO	UBE-RON
BERT BiLSTM	OGER (baseline)	0.5808	**0.6657**	0.2832	0.6838	0.8632	0.4459	0.5947	0.4581	0.5362	0.6561
	no-pretraining	0.7293	0.5939	0.7293	0.7051	0.9764	0.7932	0.9609	0.3591	0.8824	0.7076
	pretraining (Run 2a)	0.7412	0.5810	0.7455	0.7191	0.9670	0.0000	0.9584	0.3526	0.8909	0.7350
	pick-best	0.7442	0.5990	0.7483	0.7149	0.9670	0.8014*	0.9611	0.3596	**0.9027**	0.7404
	IDs (Run 1)	0.7555	0.6316	0.7966	0.7626	0.9221	0.8601	0.9669	0.4762	0.8933	0.7416
	spans+OGER	0.6586	0.6522	0.2957	0.7603	**0.9838**	0.7683	0.8451	**0.8026**	0.8163	0.6784
	IDs+spans+OGER (Run 3)	**0.7700**	0.6487	**0.8037**	**0.7645**	0.9561	**0.8705**	**0.9694**	0.5443	0.8954	**0.7488**

		CHE-BI EXT	CL EXT	GO BP EXT	GO CC EXT	GO MF EXT	MOP EXT	NCBI Taxon EXT	PR EXT	SO EXT	UBE-RON EXT
BERT BiLSTM	OGER (baseline)	0.6797	0.7236	0.3644	0.8220	0.6731	0.4106	0.5994	0.4826	0.4604	0.6810
	no-pretraining	0.8031	0.7263	0.7758	0.8674	0.7154	0.7996	0.9583	0.4004	0.9092	0.6900
	pretraining (Run 2a)	0.8173	0.7199	0.7712	0.8725	0.7411	0.5630	0.9554	0.4005	0.9167	0.7312
	pick-best	0.8168	0.7289	0.7755	0.8723	**0.7438**	0.7996*	0.9549	0.4122	**0.9187**	0.7458
	IDs (Run 1)	0.8143	0.7375	0.8085	0.8918	0.6530	0.8240	0.9682	0.4706	0.9056	0.7654
	spans+OGER	0.7180	0.7187	0.3799	0.8862	0.6715	0.4562	0.8351	**0.8011**	0.5640	0.7029
	IDs+spans+OGER	**0.8209**	**0.7484**	**0.8138**	**0.8936**	0.6691	**0.8437**	**0.9722**	0.5516	0.9069	**0.7714**

*ontology pretraining disabled

Table 1: F-Score results of our experiments using the CRAFT corpus. Underlined numbers denote submitted results; other results were obtained in post-submission experiments. Bold figures mark the best result for each entity type (column).

ing problem. This approach has been adopted by many (Zhang et al., 2014; Cho et al., 2017), also using different neural architectures (Li et al., 2017; Liu and Xu, 2018; Tutubalina et al., 2018).

There have been continued efforts to jointly address NER and NEN, fighting the problem of error propagation inherent to pipeline architectures. Dictionary-based approaches can detect and normalize concept mentions in a single step (Tseytlin et al., 2016; Pafilis et al., 2013), even though postfiltering (Basaldella et al., 2017; Cuzzola et al., 2017) or other strategies are usually required to achieve good performance. Example-based approaches include probabilistic (Leaman and Lu, 2016) and graphical (Lou et al., 2017; ter Horst et al., 2017) systems for jointly learning NER+NEN in shared or interdependent models. Zhao et al. (2019) propose a multi-task-learning set-up for neural NER and NEN with bidirectional feedback, as mentioned earlier.

Recently, it has been shown that BERT models that are pretrained on biomedical and clinical datasets are beneficial for the NER task in the biomedical domain (Lee et al., 2019; Beltagy et al., 2019). To address the NEN task with BERT-based models, Kim et al. (2019) combined the BioBERT model with a rule-based approach to multi-type resolution and a dictionary lookup for the normalization.

4 Results

The results of our experiments are summarized in Tables 1 and 2. The tables contain both officially submitted results (printed with underline) and post-submission runs. The results were obtained by the official evaluation suite, which measures performance in terms of Slot Error Rate (SER) (Makhoul et al., 1999) and F-Score (F1). Both metrics are based on the counts of matches (true positives), insertions (false positives), deletions (false negatives) and substitutions (partial positives). The substitutions, as defined by Bossy et al. (2013), are a way to give partial credit to system predictions that are partially correct, e g. when the correct ID was assigned to one token of a multi-word expression. While F1 is a measure of accurateness ranging from 1 (perfect) to 0 (no matching prediction at all), SER is a measure of errors ranging from 0 (perfect) to above 1 (more errors than ground-truth annotations). The rankings produced by the two metrics are not guaranteed to be identical; in fact, we report several cases where F1 and SER disagree on the question of which system performed best. For both metrics, the scores are micro-averaged across all 30 documents of the test set.

We used the plain dictionary-based system OGER as a baseline. For the BiLSTM system, we compared three different configurations: no-

		CHE-BI	CL	GO BP	GO CC	GO MF	MOP	NCBI Taxon	PR	SO	UBE-RON
BERT BiLSTM	OGER (baseline)	0.7873	**0.4862**	0.9826	0.6120	0.3032	1.6238	1.0122	1.9768	1.2617	0.5584
	no-pretraining	0.4280	0.5628	0.4231	0.4829	0.0443	0.3507	0.0688	0.9017	0.1934	0.4379
	pretraining (Run 2a)	0.4089	0.5781	0.3991	0.4296	0.0638	1.0000	0.0733	0.8597	0.1786	0.3913
	pick-best	0.4038	0.5563	0.3956	0.4455	0.0638	0.3453*	0.0723	0.8501	**0.1593**	0.3864
	IDs (Run 1)	0.3569	0.5780	0.3145	0.3848	0.1507	0.2882	0.0580	0.7612	0.1689	**0.3752**
	spans+OGER	0.5111	0.5000	0.8276	**0.3788**	**0.0319**	0.3762	0.2240	**0.3052**	0.2918	0.4749
	IDs+spans+OGER (Run 3)	**0.3388**	0.5620	**0.3047**	0.3888	0.0869	**0.2684**	**0.0537**	0.6863	0.1680	0.3770

		CHE-BI EXT	CL EXT	GO BP EXT	GO CC EXT	GO MF EXT	MOP EXT	NCBI Taxon EXT	PR EXT	SO EXT	UBE-RON EXT
BERT BiLSTM	OGER (baseline)	0.6032	**0.3361**	0.8677	0.3493	0.5459	1.8108	0.9869	1.7056	1.1596	0.5210
	no-pretraining	0.3152	0.3555	0.3398	0.2076	0.4266	0.3445	0.0744	0.8354	0.1398	0.4552
	pretraining (Run 2a)	0.3016	0.3547	0.3357	0.2032	0.3922	0.5564	0.0776	0.8047	0.1257	0.3943
	pick-best	0.3016	0.3497	0.3333	0.2051	**0.3881**	0.3445*	0.0784	0.7715	**0.1230**	0.3730
	IDs (Run 1)	0.2664	0.3667	0.2867	**0.1678**	0.5081	0.3440	0.0538	0.7257	0.1475	**0.3371**
	spans+OGER	0.4224	0.3417	0.7419	0.1907	0.4676	0.6432	0.2353	**0.3030**	0.5566	0.4450
	IDs+spans+OGER	**0.2571**	0.3583	**0.2786**	0.1681	0.4999	**0.3080**	**0.0466**	0.6464	0.1466	0.3384

*ontology pretraining disabled

Table 2: SER results of our experiments using the CRAFT corpus. For mark-up (underline/bold) see Table 1.

pretraining, pretraining, and pick-best. For the no-pretraining run, we skipped the pretraining phase over the ontology names. The pretraining run corresponds to the description in Section 3.2; we (unofficially[3]) submitted this run as Run 2a, except for MOP and MOP_EXT, where pretraining was disabled since it had an extraordinarily negative effect for this entity type in early experiments already. In the pick-best run, we trained each model two or three times and picked the one with the best performance on the held-out set in the cross-validation; again, ontology pretraining was disabled for MOP[_EXT] for this run.

For the transformer architecture, we also compared three systems: BERT-IDs, BERT-spans+OGER, and BERT-IDs+BERT-spans+OGER. BERT-IDs was trained to predict concept identifiers directly; we submitted these results as Run 1 (except for CL_EXT, GO_CC_EXT, MOP_EXT, NCBITaxon_EXT, and UBERON_EXT, which we analyzed only in post-submission experiments due to time constraints). BERT-spans+OGER combines IOBES predictions with annotations from OGER in a pipeline fashion. The last configuration combines the previous two in a backoff manner; this was submitted as Run 3 (extension types post-submission only).

For many entity types, the BERT systems beat the BiLSTM systems, which in turn clearly outperformed the dictionary-based baseline. A notable exception to this pattern is CL, where no neural system was as accurate as OGER. However, the baseline is beaten by all other systems in many cases; this is particularly true for SER, where the baseline shows very poor performance for a number of entity types.

Among the BiLSTM systems, the effect of ontology pretraining is somewhat heterogeneous; while it clearly improved performance for some entity types (such as CHEBI[_EXT], UBERON[_EXT]), it had a marginal or even negative effect on others (e. g. NCBITaxon[_EXT]). As expected from the cross-validation results, ontology pretraining heavily decreased performance for MOP and MOP_EXT. The pick-best setting yielded modest improvements in most of the cases. In three cases (GO_MF_EXT, SO, SO_EXT), this configuration achieves the best overall scores.

Among the BERT-based systems, directly predicting IDs usually gave better results than joining span predictions with OGER annotations, and combining the two systems in a backoff manner yielded another improvement. However, the span detector coupled with OGER outperformed the two ID taggers in five cases (CL, GO_MF[_EXT], PR[_EXT]), three of which constitute best overall scores (GO_MF, PR[_EXT]). The most notable results are the ones for PR and PR_EXT, where BERT-spans+OGER beat all other systems by a margin of more than 0.25 F1/0.34 SER.

[3] after the deadline, but before the release of the ground-truth annotations

	ground-truth concepts		OGER		BiLSTM pretraining		BiLSTM pick-best		BERT-spans+ OGER		BERT-IDs+ BERT-spans+ OGER	
	unique	occ.	P	R	P	R	P	R	P	R	P	R
CHEBI	110	447	0.33	0.65	–		–		0.74	0.47	0.70	0.11
CHEBI_EXT	134	538	0.37	0.71	–		–		0.62	0.49	0.76	0.09
CL	52	484	0.72	0.31	–		–		0.88	0.22	0.59	0.04
CL_EXT	52	484	0.72	0.31	–		–		0.71	0.25	0.71	0.11
GO_BP	120	484	0.21	0.25	–		–		0.56	0.12	0.66	0.06
GO_BP_EXT	126	508	0.22	0.28	–		–		0.29	0.18	0.62	0.07
GO_CC	32	184	0.19	0.35	–		–		0.50	0.17	0.49	0.06
GO_CC_EXT	36	231	0.28	0.47	–		–		0.58	0.19	0.60	0.07
GO_MF	1	1	0.10	0.50	–		–		–		–	
GO_MF_EXT	73	416	0.38	0.22	–		–		0.57	0.15	0.54	0.04
MOP	2	2	0.08	1.00	–		–		–		–	
MOP_EXT	2	2	0.08	1.00	–		–		–		–	
NCBITaxon	40	87	0.02	0.50	–		–		0.40	0.34	0.75	0.22
NCBITaxon_EXT	44	95	0.02	0.54	–		–		0.43	0.35	0.85	0.25
PR	278	4782	0.26	0.86	0.61	4E-4	0.63	4E-4	0.81	0.74	0.69	0.15
PR_EXT	309	5156	0.27	0.84	0.22	3E-3	0.34	8E-3	0.84	0.73	0.65	0.20
SO	16	101	0.04	0.87	–		–		0.10	0.06	0.52	0.02
SO_EXT	25	123	0.05	0.78	–		–		0.28	0.47	0.85	0.41
UBERON	203	1297	0.47	0.33	0.74	2E-3	0.69	2E-3	0.74	0.25	0.59	0.06
UBERON_EXT	207	1308	0.47	0.33	0.76	2E-3	0.87	1E-3	0.78	0.27	0.60	0.06

Table 3: System performance for unseen concepts: precision (P) and recall (R) calculated over the subset of annotations and predictions of IDs that were absent from the training data. A dash (–) denotes that the system only predicted known IDs for the given entity type. The systems BiLSTM no-pretraining and BERT-IDs are omitted as they cannot predict unseen labels.

5 Discussion

The results show that, in general, neural sequence taggers can be successfully applied to biomedical concept recognition, using a single model for joint NER+NEN. Unfortunately, we cannot compare our results to other work, as no other team has submitted results to the concept-annotation task and no official baseline is available at the time of writing. Since the CRAFT test set has only been released in the course of the present shared task, it is not possible to directly benchmark our results against previous work (such as Funk et al., 2014; Tseytlin et al., 2016; Hailu, 2019) either. However, the tested systems allow for a comparison of different approaches.

The strategies for extending the concept coverage – a vital feature for many applications – show a mixed picture. Pretraining on ontology names has led to limited benefit only. While it has demonstrated a positive effect for many entity types, it has been able to increase the set of recognized concepts only occasionally. As can be seen in Table 3, ontology pretraining led to prediction of IDs outside the training data in four entity types (PR[_EXT], UBERON[_EXT]). Even though the majority of the predicted unseen IDs is correct, they only account for a fraction of the ground-truth annotations.

On the other hand, combining BERT span predictions with OGER annotations resulted in correct predictions of unseen IDs for almost all entity types – the exceptions being GO_MF, MOP, and MOP_EXT, which suffer from a small number of concepts or positive examples in the training data. The BERT-spans+OGER system is particularly strong for PR[_EXT], where recognizing unseen concepts is especially important due to the diversity and abundance of protein mentions in the literature. When this system is used as a backoff for BERT-IDs, the recall for unseen concepts drops due to the bias for existing knowledge inherent to the ID tagger. In some cases this bias is beneficial for precision, i. e. the ID tagger suppresses many false-positive predictions of OGER (e. g. CHEBI_EXT, NCBITaxon[_EXT], SO[_EXT]), while in other cases false positives of the ID tagger hide correct OGER predictions, leading to lower precision.

A few examples of correctly predicted IDs absent from the training corpus are given in context in the following. BERT-IDs+BERT-spans+OGER predicted CHEBI_PR_EXT:somatostatin in document 17503968 (two occurrences):

> However, the **somatostatin receptor 2** (**SSTR-2**) antagonist PRL-2903 does

> not interfere with the ability of glucose (at 3 and 7 mM) to inhibit glucagon secretion from mouse islets [47].

The same system predicted CHEBI:60004 in document 11604102:

> Adult mouse testes were homogenized in a buffer containing 20 mM Tris, pH 7.5, 100 mM KCl, 5 mM MgCl2, 0.3% NP-40, 40 U/ml of Rnasin ribonuclase inhibitor (Promega, Madison, WI), and a **mixture** of 10 protease inhibitors provided [...]

BiLSTM pick-best predicted PR:000008373 in document 16968134:

> Decreased Osteogenic Differentiation Correlates with Abnormal Distribution of **Cx43**

The creators of the CRAFT corpus have put great effort in building an annotated corpus with high quality and consistency across all entity types. However, the diversity of the different types requires a lot of engineering for tackling them all. A single approach is not sufficient to meet the differing needs of all entity types. The experiments with the test set have yielded a few surprising results, such as the comparatively good performance of the dictionary-based approach on CL or the outstanding scores for BERT-spans+OGER on PR[_EXT].

Of the two concept extension strategies, the NER+dictionary backoff has worked well, whereas the effect of ontology pretraining was not too conclusive. Since we tested each of the strategies with only one system architecture, it is not entirely clear which component contributed the most to the success – the network architecture or the extension strategy. Testing the inverse combinations, i. e. BERT with ontology pretraining and BiLSTM with OGER backoff, is left for future work.

Acknowledgments

The research activities of the OntoGene/BioMeXT group at the University of Zurich are supported by the Swiss National Science Foundation (grant CR30I1_162758). We would like to thank the organisers for this well-organised shared task with high-quality annotations and prompt support. Additional thanks to the anonymous reviewers who provided valuable feedback.

References

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):1–20.

Marco Basaldella, Lenz Furrer, Carlo Tasso, and Fabio Rinaldi. 2017. Entity recognition in the biomedical domain using a hybrid approach. *Journal of Biomedical Semantics*, 8(1):51.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Philippe Bessières, and Claire Nédellec. 2013. BioNLP shared task 2013 – an overview of the bacteria biotope task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 161–169, Sofia, Bulgaria. Association for Computational Linguistics.

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14(281).

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany. Association for Computational Linguistics.

Hyejin Cho, Wonjun Choi, and Hyunju Lee. 2017. A method for named entity normalization in biomedical articles: application to diseases and plants. *BMC Bioinformatics*, 18(1):451.

K. Bretonnel Cohen, Karin Verspoor, Karën Fort, Christopher Funk, Michael Bada, Martha Palmer, and Lawrence E. Hunter. 2017. *The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain*, pages 1379–1394. Springer Netherlands, Dordrecht.

John Cuzzola, Jelena Jovanović, and Ebrahim Bagheri. 2017. RysannMD: a biomedical semantic annotator balancing speed and accuracy. *Journal of Biomedical Informatics*, 71:91–109.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, MN, USA.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 297–302, Beijing, China. Association for Computational Linguistics.

Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K. Bretonnel Cohen, Lawrence E. Hunter, and Karin Verspoor. 2014. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1):1–29.

Lenz Furrer, Anna Jancso, Nicola Colic, and Fabio Rinaldi. 2019. OGER++: hybrid multi-type entity recognition. *Journal of Cheminformatics*, 11(1):7.

Omid Ghiasvand and Rohit J. Kate. 2014. UWM: disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 828–832, Dublin, Ireland. Association for Computational Linguistics.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Negacy Degefa Hailu. 2019. *Investigation of traditional and deep neural sequence models for biomedical concept recognition*. Ph.D. thesis, University of Colorado at Denver, Anschutz Medical Campus.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hendrik ter Horst, Matthias Hartung, and Philipp Cimiano. 2017. *Joint Entity Recognition and Linking in Technical Domains Using Undirected Probabilistic Graphical Models*, volume 10318, pages 166–180. Springer.

Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, Minneapolis, MN, USA. Association for Computational Linguistics.

Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access*, 7:73729–73740.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839.

Hsin-Chun Lee, Yi-Yu Hsu, and Hung-Yu Kao. 2016. AuDis: an automatic CRF-enhanced disease normalization in biomedical text. *Database*, 2016:baw091.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. Btz682.

Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. CNN-based ranking for biomedical entity normalization. *BMC Bioinformatics*, 18(11):385.

Hongwei Liu and Yun Xu. 2018. A deep learning way for disease name representation and normalization. In Xuanjing Huang, Jing Jiang, Dongyan Zhao, Yansong Feng, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 151–157. Springer International Publishing.

Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics*, 33(15):2363–2371.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252.

Alejandro Metke-Jimenez and Sarvnaz Karimi. 2016. Concept identification and normalisation for adverse drug event discovery in medical forums. In *Proceedings of the First International Workshop on Biomedical Data Integration and Discovery (BMDID 2016)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119. Curran Associates, Inc.

Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLOS ONE*, 8(6):1–6.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Sunghwan Sohn, Donald C. Comeau, Won Kim, and W. John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, 9(1):402.

Eugene Tseytlin, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan, and Rebecca S. Jacobson. 2016. NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*, 17(1):1–15.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84:93–102.

Yaoyun Zhang, Jingqi Wang, Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, and Hua Xu. 2014. UTH_CCB: a report for SemEval 2014 – Task 7 analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 802–806, Dublin, Ireland. Association for Computational Linguistics.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 817–824.